NGS Galaxy Exercise

February 20th 2012

Agata, Arcadio and Kirstine

Use an internet cable, no wireless!

Login to Galaxy installed on a CBS server called kampen

Go to a internet browser and write: http://kampen

Login with your username and password. Handed out.

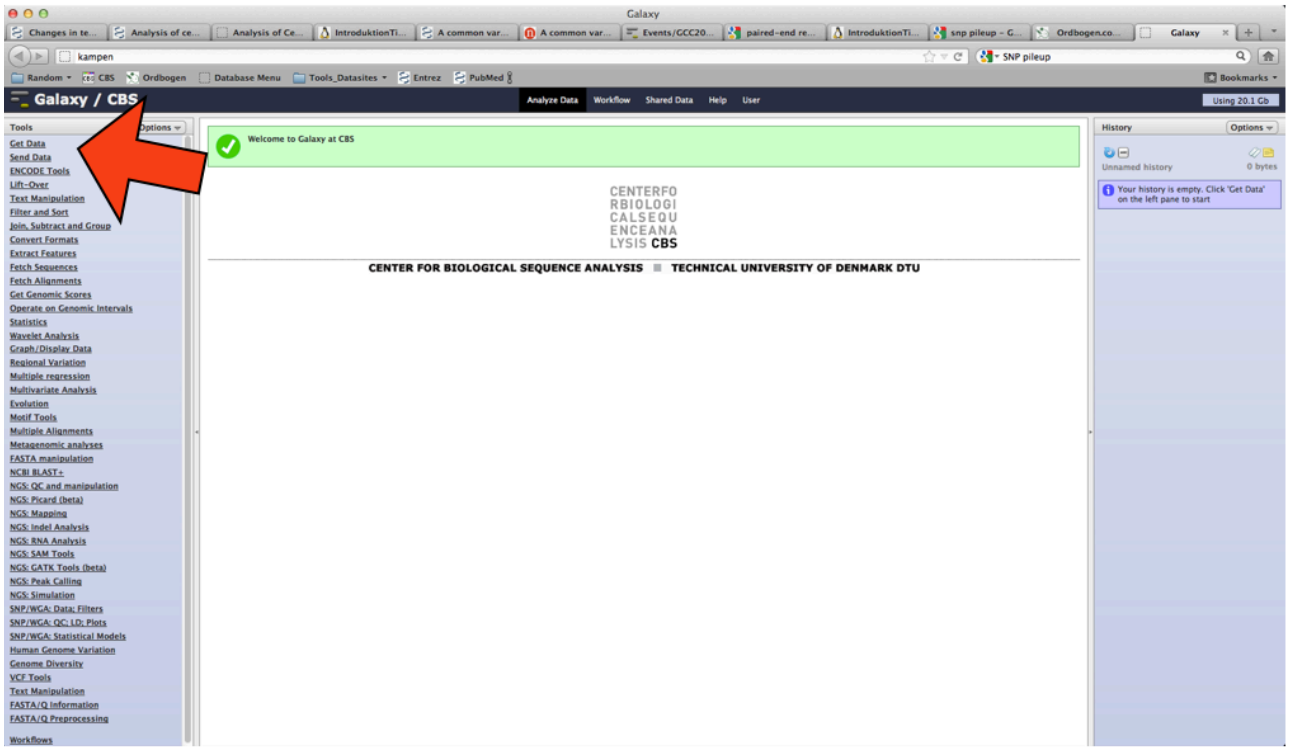
Data files

Breast cancer cell line, chromosome 13

Illumina reads

The raw reads has been trimmed (remove or trim bad quality reads)

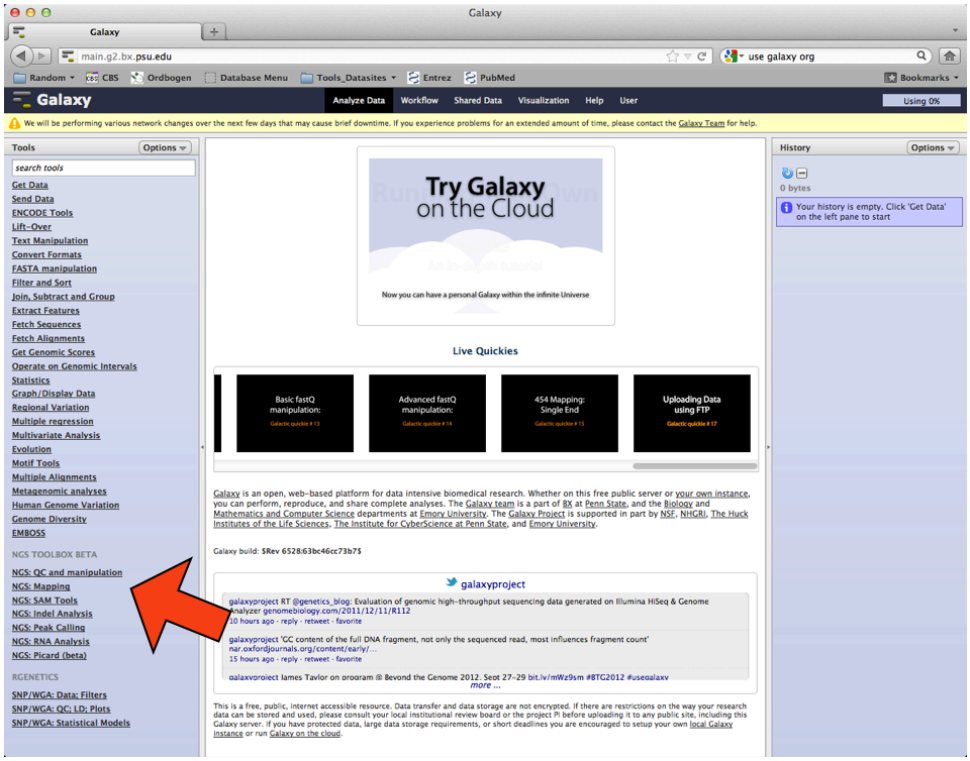
Paired-end reads – two files



* Get Data --> Upload file
File format: fastqillumina
Import files via URL:
http://www.cbs.dtu.dk/services/HumLoc-1.0/trimmed1.fq
http://www.cbs.dtu.dk/services/HumLoc-1.0/trimmed2.fq

Genome: Human Feb. 2009 (GRCh37/hg19) (hg19)

Quality control on both Fastq files
* NGS: QC and manipulation --> Fastqc: Fastqc QC



Look at the quality report. Click at the eye next to the FastQC.html report

Question A: If the reads on an accessible read quality? Above phred score 20

Align reads to reference genome (chromosome 13)
Use Bowtie to align the paired end reads
* NGS: Mapping ---> Map with Bowtie for Illumina
  Reference genome: GRCh37/hg19
  Mate-paired: Paired-end
  Forward fastq file: "..._1.fq"
  Reverse fatsq file: "..._2.fq"
  BWA settings: Common

Output of alignment is a BAM file

Question B: What is the difference between single-end and paired-end reads?

Convert SAM file to BAM file
* NGS: SAM tools -> SAM-to-BAM

Check how many reads in the file is aligned to the reference genome
* NGS: SAM tool --> flagstat
  Use the BAM files

Question C: How many percentage of the reads are mapped to the reference genome?

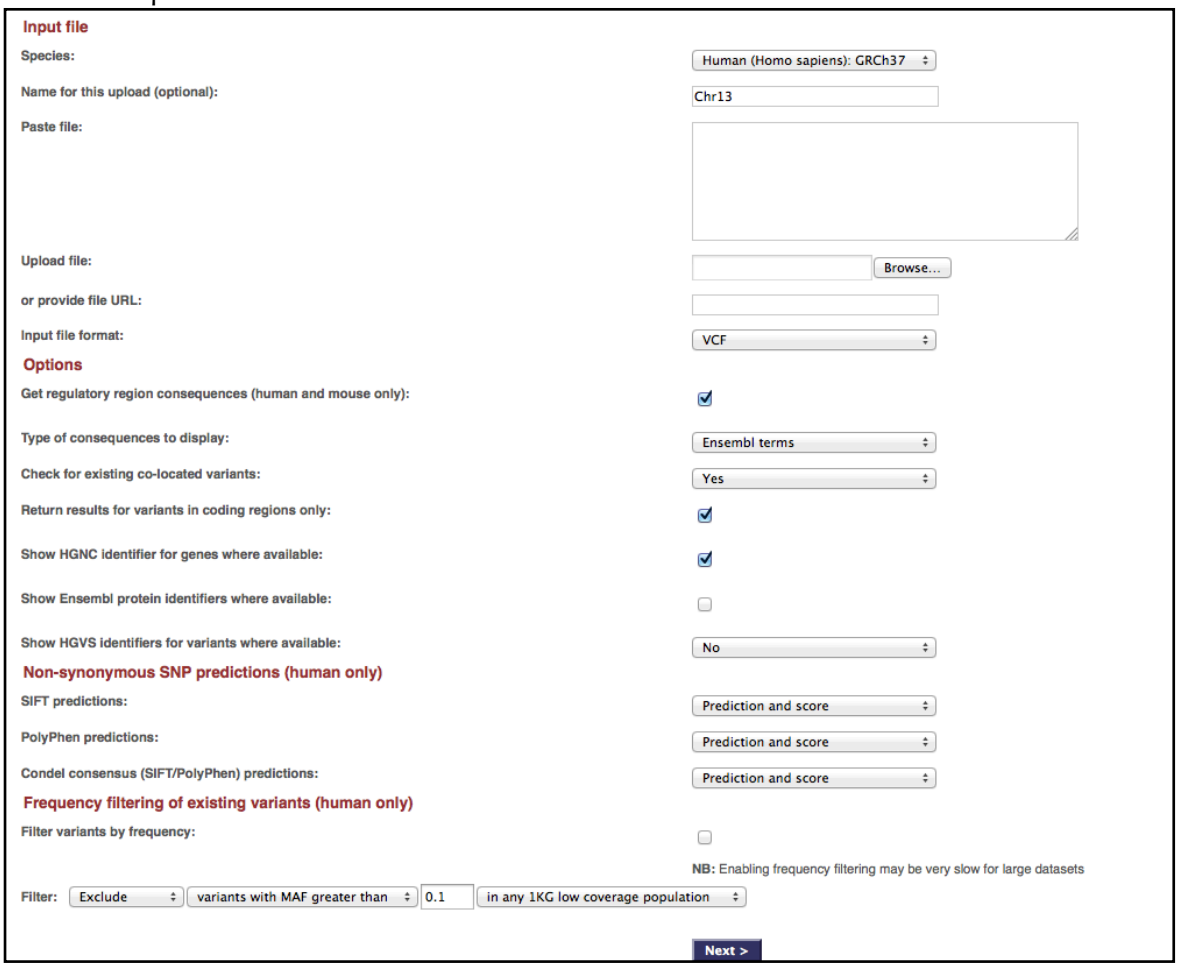
Remove PCR duplicates
* NGS: SAM Tools --> rmdup

Generate pileup of SNPs
Identify all the SNPs detected with alignment
* NGS: SAM Tools --> MPileup
  Select the BAM file without duplicates
  Genotype Likelihood Computation: Perform genotype likelihood computation
  Set advanced options:
    - Minimum mapping quality = 30
    - Minimum base quality = 20

The MPileup output is in bcf format (binary). We have converted the file to vcf file and in the same go filtered by depth 30 (the minimum number of reads covering the SNPs).
The file is available via this link: http://cbs.dtu.dk/services/HumLoc-1.0/chr13.flt.vcf

Question D: All the reads were supposed to map to chromosome 13. Since you see "high quality" SNPs on other chromosomes what can you say about the quality of the alignment?

Investigate SNPs in Ensembl
Upload the file to Ensembl: http://www.ensembl.org/Homo_sapiens/UserData/UploadVariations
Upload file
Choose the parameters like the screenshot below



Question E: Find BRCA2. How many non-synonymous coding mutations are detected in BRCA2?

Question F: How many of the BRCA2 SNPs are predicted to be deleterious?

Question F: Can you identify the mutation from the article? If so, is the mutation predicted to be deleterious?